DEPARTMENT OF
POLITICS
AND
INTERNATIONAL
RELATIONS

Lecture Notes (corrected)

Introduction to Nonparametric Regression

John Fox

McMaster University
Canada

1. What is Nonparametric Regression?

Regression analysis traces the average value of a response variable (y) as a function of one or several predictors (x's).

Suppose that there are two predictors, x_1 and x_2.

- The object of regression analysis is to estimate the *population regression function* $\mu|x_1, x_2 = f(x_1, x_2)$.
- Alternatively, we may focus on some other aspect of the conditional distribution of y given the x's, such as the median value of y or its variance.

As it is usually practiced, regression analysis assumes:

- a linear relationship of y to the x's, so that
$$\mu|x_1, x_2 = f(x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2$$
- that the conditional distribution of y is, except for its mean, everywhere the same, and that this distribution is a normal distribution
$$y \sim N(\alpha + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$$
- that observations are sampled independently, so the y_i and $y_{i'}$ are independent for $i \neq i'$.
- The full suite of assumptions leads to linear least-squares regression.

These are strong assumptions, and there are many ways in which they can go wrong. For example:

- as is typically the case in time-series data, the errors may not be independent;
- the conditional variance of y (the 'error variance') may not be constant;
- the conditional distribution of y may be very non-normal — heavy-tailed or skewed.

Nonparametric regression analysis relaxes the assumption of linearity, substituting the much weaker assumption of a smooth population regression function $f(x_1, x_2)$.

- The cost of relaxing the assumption of linearity is much greater computation and, in some instances, a more difficult-to-understand result.
- The gain is potentially a more accurate estimate of the regression function.

Some might object to the 'atheoretical' character of nonparametric regression, which does not specify the form of the regression function $f(x_1, x_2)$ in advance of examination of the data. I believe that this objection is ill-considered:

- Social theory might suggest that y depends on x_1 and x_2, but it is unlikely to tell us that the relationship is linear.
- A necessary condition of effective statistical data analysis is for statistical models to summarize the data accurately.

In this short-course, I will first describe nonparametric *simple* regression, where there is a quantitative response variable y and a single predictor x, so $y = f(x) + \varepsilon$.

I'll then proceed to nonparametric *multiple* regression — where there are several predictors, and to *generalized nonparametric regression* models — for example, for a dichotomous (two-category) response variable.

The course is based on materials from Fox, *Nonparametric Simple Regression*, and Fox, *Multiple and Generalized Nonparametric Regression* (both Sage, 2000).

Starred (*) sections will be covered time permitting.

2. Preliminary Examples

2.1 Infant Mortality

Figure 1 (a) shows the relationship between infant-mortality rates (infant deaths per 1,000 live births) and GDP per capita (in U. S. dollars) for 193 nations of the world.

- The nonparametric regression line on the graph was produced by a method called *lowess* (or *loess*), an implementation of local polynomial regression, and the most commonly available method of nonparametric regression.
- Although infant mortality declines with GDP, the relationship between the two variables is highly nonlinear: As GDP increases, infant mortality initially drops steeply, before leveling out at higher levels of GDP.

Because both infant mortality and GDP are highly skewed, most of the data congregate in the lower-left corner of the plot, making it difficult to discern the relationship between the two variables. The linear least-squares fit to the data does a poor job of describing this relationship.

- In Figure 1 (b), both infant mortality and GDP are transformed by taking logs. Now the relationship between the two variables is nearly linear.

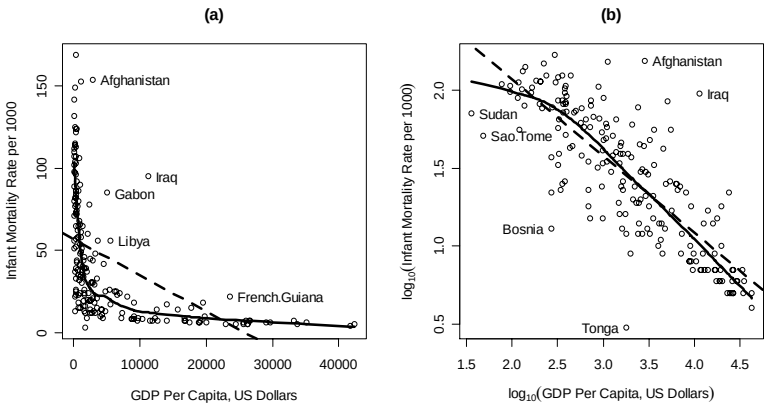

Figure 1. Infant-mortality rate per 1000 and GDP per capita (US dollars) for 193 nations.

2.2 Women's Labour-Force Participation

An important application of generalized nonparametric regression is to binary data. Figure 2 shows the relationship between married women's labour-force participation and the log of the women's 'expected wage rate.'

- The data, from the 1976 U. S. Panel Study of Income Dynamics were originally employed by Mroz (1987), and were used by Berndt (1991) as an exercise in linear logistic regression and by Long (1997) to illustrate that method.
- Because the response variable takes on only two values, I have vertically 'jittered' the points in the scatterplot.
- The nonparametric logistic-regression line shown on the plot reveals the relationship to be curvilinear. The linear logistic-regression fit, also shown, is misleading.

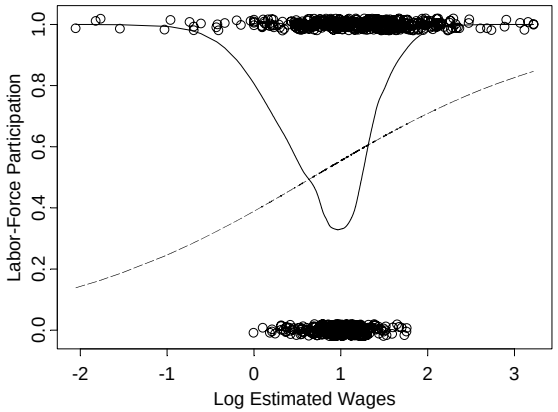

Figure 2. Scatterplot of labor-force participation (1 = Yes, 0 = No) by the log of estimated wages.

2.3 Occupational Prestige

Blishen and McRoberts (1976) reported a linear multiple regression of the rated prestige of 102 Canadian occupations on the income and education levels of these occupations in the 1971 Canadian census. The purpose of this regression was to produce substitute predicated prestige scores for many other occupations for which income and education levels were known, but for which direct prestige ratings were unavailable.

- Figure 3 shows the results of fitting an *additive nonparametric regression* to Blishen's data:

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon$$

- The graphs in Figure 3 show the estimated partial regression functions for income $\widehat{f_1}$ and education $\widehat{f_2}$. The function for income is quite nonlinear, that for education somewhat less so.

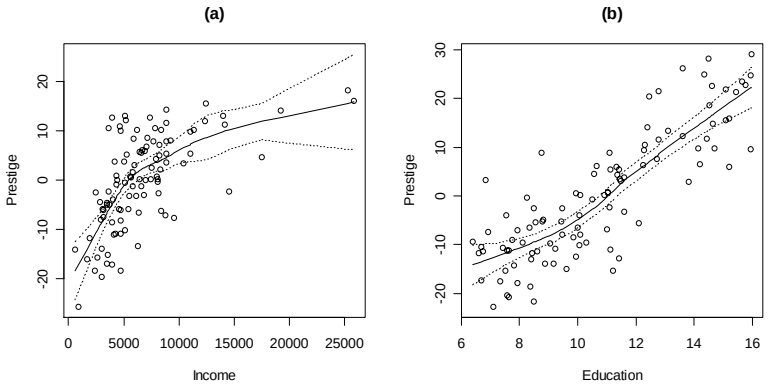

Figure 3. Plots of the estimated partial-regression functions for the additive regression of prestige on the income and education levels of 102 occupations.

3. Nonparametric Simple Regression

Most interesting applications of regression analysis employ several predictors, but nonparametric simple regression is nevertheless useful for two reasons:

1. Nonparametric simple regression is called *scatterplot smoothing*, because the method passes a smooth curve through the points in a scatterplot of y against x. Scatterplots are (or should be!) omnipresent in statistical data analysis and presentation.
2. Nonparametric simple regression forms the basis, by extension, for nonparametric multiple regression, and directly supplies the building blocks for a particular kind of nonparametric multiple regression called *additive regression*.

3.1 Binning and Local Averaging

Suppose that the predictor variable x is discrete (e.g., x is age at last birthday and y is income in dollars). We want to know how the average value of y (or some other characteristics of the distribution of y) changes with x; that is, we want to know $\mu|x$ for each value of x.

- Given data on the entire population, we can calculate these conditional population means directly.
- If we have a very large sample, then we can calculate the sample average income for each value of age, $\overline{y}|x$; the estimates $\overline{y}|x$ will be close to the population means $\mu|x$.

Figure 4 shows the median and quartiles of the distribution of income from wages and salaries as a function of single years of age. The data are taken from the 1990 U. S. Census one-percent Public Use Microdata Sample, and represent 1.24 million observations.

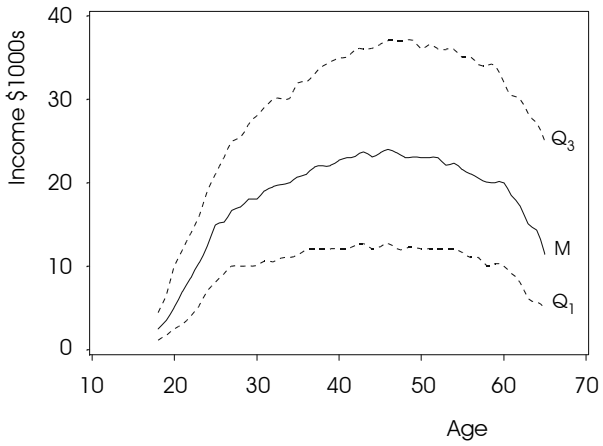

Figure 4. Simple nonparametric regression of income on age, with data from the 1990 U. S. Census one-percent sample.

3.1.1 Binning

Now suppose that the predictor variable x is continuous. Instead of age at last birthday, we have each individual's age to the minute.

- Even in a very large sample, there will be very few individuals of precisely the same age, and conditional sample averages $\overline{y}|x$ would therefore each be based on only one or a few observations.
- Consequently, these averages will be highly variable, and will be poor estimates of the population means $\mu|x$.

Because we have a very large sample, however, we can dissect the range of x into a large number of narrow class intervals or *bins*.

- Each bin, for example, could constitute age rounded to the nearest year (returning us to single years of age). Let $x_1, x_2, ..., x_b$ represent the x-values at the bin centers.
- Each bin contains a lot of data, and, consequently, the conditional sample averages, $\overline{y}_i = \overline{y}|(x \text{ in bin } i)$, are very stable.
- Because each bin is narrow, these bin averages do a good job of estimating the regression function $\mu|x$ anywhere in the bin, including at its center.

Given sufficient data, there is essentially no cost to binning, but in smaller samples it is not practical to dissect the range of x into a large number of narrow bins:

- There will be few observations in each bin, making the sample bin averages $\overline{y}_i$ unstable.
- To calculate stable averages, we need to use a relatively small number of wider bins, producing a cruder estimate of the population regression function.

There are two obvious ways to proceed:

1. We could dissect the range of x into bins of equal width. This option is attractive only if x is sufficiently uniformly distributed to produce stable bin averages based on a sufficiently large number of observations.
2. We could dissect the range of x into bins containing roughly equal numbers of observations.

Figure 5 depicts the binning estimator applied to the U. N. infant-mortality data. The line in this graph employs 10 bins, each with roughly 19 observations.

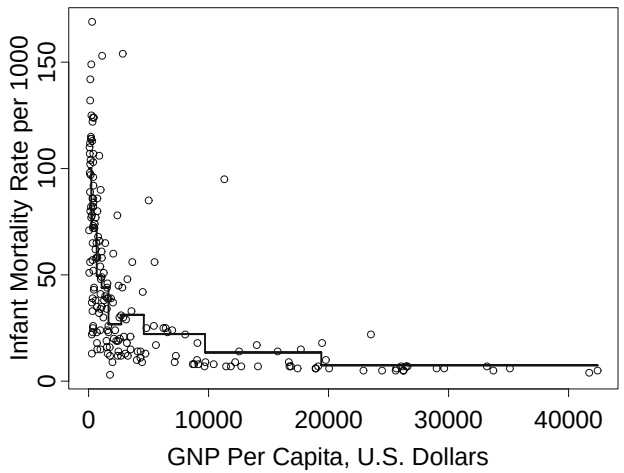

Figure 5. The binning estimator applied to the relationship between infant mortality and GDP per capita.

Treating a discrete quantitative predictor variable as a set of categories and binning continuous predictor variables are common strategies in the analysis of large datasets.

- Often continuous variables are implicitly binned in the process of data collection, as in a sample survey that asks respondents to report income in class intervals (e.g., \$0–\$5000, \$5000–\$10,000, \$10,000–\$15,000, etc.).
- If there are sufficient data to produce precise estimates, then using dummy variables for the values of a discrete predictor or for the class intervals of a binned predictor is preferable to blindly assuming linearity.
- An even better solution is to compare the linear and nonlinear specifications.

3.1.2 Statistical Considerations*

The mean-squared error of estimation is the sum of squared bias and sampling variance:

$$\text{MSE}[\widehat{f}(x_0)] = \text{bias}^2[\widehat{f}(x_0)] + V[\widehat{f}(x_0)]$$

As is frequently the case in statistical estimation, minimizing bias and minimizing variance work at cross purposes:

- Wide bins produce small variance and large bias.
- Small bins produce large variance and small bias.
- Only if we have a very large sample can we have our cake and eat it too.
- All methods of nonparametric regression bump up against this problem in one form or another.

Even though the binning estimator is biased, it is *consistent* as long as the population regression function is reasonably smooth.

- All we need do is shrink the bin width to 0 as the sample size n grows, but shrink it sufficiently slowly that the number of observations in each bin grows as well.
- Under these circumstances, $\text{bias}[\widehat{f}(x)] \to 0$ and $V[\widehat{f}(x)] \to 0$ as $n \to \infty$.

3.1.3 Local Averaging

The essential idea behind *local averaging* is that, as long as the regression function is smooth, observations with x-values near a focal x_0 are informative about $f(x_0)$.

- Local averaging is very much like binning, except that rather than dissecting the data into non-overlapping bins, we move a bin (called a *window*) continuously over the data, averaging the observations that fall in the window.
- We can calculate $\widehat{f}(x)$ at a number of focal values of x, usually equally spread within the range of observed x-values, or at the (ordered) observations, $x_{(1)}, x_{(2)}, ..., x_{(n)}$.
- As in binning, we can employ a window of fixed width w centered on the focal value x_0, or can adjust the width of the window to include a constant number of observations, m. These are the m *nearest neighbors* of the focal value.

- Problems occur near the extremes of the x's. For example, all of the nearest neighbors of $x_{(1)}$ are greater than or equal to $x_{(1)}$, and the nearest neighbors of $x_{(2)}$ are almost surely the same as those of $x_{(1)}$, producing an artificial flattening of the regression curve at the extreme left, called *boundary bias*. A similar flattening occurs at the extreme right, near $x_{(n)}$.

Figure 6 shows how local averaging works, using the relationship of prestige to income in the Canadian occupational prestige data.

1. The window shown in panel (a) includes the $m = 40$ nearest neighbors of the focal value $x_{(80)}$.
2. The y-values associated with these observations are averaged, producing the fitted value $\widehat{y}_{(80)}$ in panel (b).
3. Fitted values are calculated for each focal x (in this case $x_{(1)}, x_{(2)}, ..., x_{(102)}$) and then connected, as in panel (c).

- In addition to the obvious flattening of the regression curve at the left and right, local averages can be rough, because $\widehat{f}(x)$ tends to take small jumps as observations enter and exit the window. The kernel estimator (described shortly) produces a smoother result.
- Local averages are also subject to distortion when outliers fall in the window, a problem addressed by robust estimation.

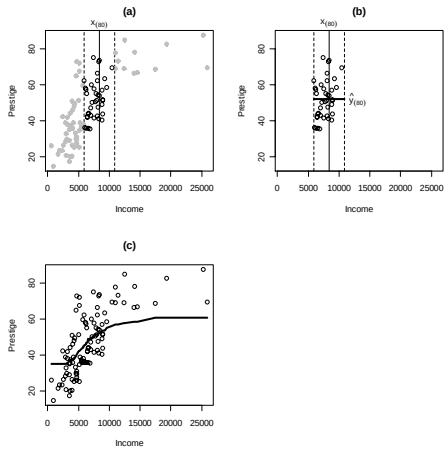

Figure 6. Nonparametric regression of prestige on income using local averages.

3.2 Kernel Estimation (Locally Weighted Averaging)

Kernel estimation is an extension of local averaging.

- The essential idea is that in estimating $f(x_0)$ it is desirable to give greater weight to observations that are close to the focal x_0.
- Let $z_i = (x_i - x_0)/h$ denote the scaled, signed distance between the x-value for the ith observation and the focal x_0. The scale factor h, called the *bandwidth* of the kernel estimator, plays a role similar to the window width of a local average.
- We need a *kernel function* $K(z)$ that attaches greatest weight to observations that are close to the focal x_0, and then falls off symmetrically and smoothly as $|z|$ grows. Given these characteristics, the specific choice of a kernel function is not critical.

- Having calculated weights $w_i = K[(x_i - x_0)/h]$, we proceed to compute a fitted value at x_0 by weighted local averaging of the y's:

$$\widehat{f}(x_0) = \widehat{y}|x_0 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

- Two popular choices of kernel functions, illustrated in Figure 7, are the *Gaussian* or *normal kernel* and the *tricube kernel*:
  - The normal kernel is simply the standard normal density function,

$$K_N(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

    Here, the bandwidth h is the standard deviation of a normal distribution centered at x_0.

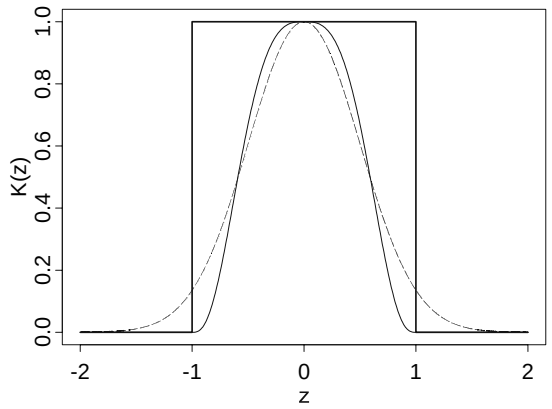

Figure 7. Tricube (light solid line), normal (broken line, rescaled) and rectangular (heavy solid line) kernel functions.

  - The tricube kernel is

$$K_T(z) = \begin{cases} (1 - |z|^3)^3 & \text{for } |z| < 1 \\ 0 & \text{for } |z| \geq 1 \end{cases}$$

    For the tricube kernel, h is the half-width of a window centered at the focal x_0. Observations that fall outside of the window receive 0 weight.
  - Using a *rectangular kernel* (also shown in Figure 7)

$$K_R(z) = \begin{cases} 1 & \text{for } |z| < 1 \\ 0 & \text{for } |z| \geq 1 \end{cases}$$

    gives equal weight to each observation in a window of half-width h centered at x_0, and therefore produces an *unweighted* local average.

I have implicitly assumed that the bandwidth h is fixed, but the kernel estimator is easily adapted to nearest-neighbour bandwidths.

- The adaptation is simplest for kernel functions, like the tricube kernel, that fall to 0: Simply adjust $h(x)$ so that a fixed number of observations m are included in the window.
- The fraction m/n is called the *span* of the kernel smoother.

Kernel estimation is illustrated in Figure 8 for the Canadian occupational prestige data.

- Panel (a) shows a neighborhood containing 40 observations centered on the 80th ordered x-value.
- Panel (b) shows the tricube weight function defined on the window; the bandwidth $h[x_{(80)}]$ is selected so that the window that accommodates the 40 nearest neighbors of the focal $x_{(80)}$. Thus, the span of the smoother is $40/102 \simeq .4$.
- Panel (c) shows the locally weighted average, $\widehat{y}_{(80)} = \widehat{y}|x_{(80)}$.
- Panel (d) connects the fitted values to obtain the kernel estimate of the regression of prestige on income. In comparison with the local-average regression, the kernel estimate is smoother, but it still exhibits flattening at the boundaries.

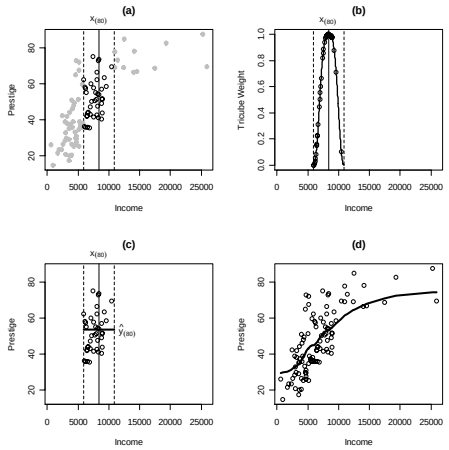

Figure 8. The kernel estimator applied to the Canadian occupational prestige data.

Varying the bandwidth of the kernel estimator controls the smoothness of the estimated regression function: Larger bandwidths produce smoother results. Choice of bandwidth will be discussed in more detail in connection with local polynomial regression.

3.3 Local Polynomial Regression

Local polynomial regression corrects some of the deficiencies of kernel estimation.

- It provides a generally adequate method of nonparametric regression that extends to multiple regression, additive regression, and generalized nonparametric regression.
- An implementation of local polynomial regression called *lowess* (or *loess*) is the most commonly available method of nonparametric regression.

Perhaps you are familiar with polynomial regression, where a p-degree polynomial in a predictor x,

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

is fit to data, usually by the method of least squares:

- $p = 1$ corresponds to a linear fit, $p = 2$ to a quadratic fit, and so on.
- Fitting a constant (i.e., the mean) corresponds to $p = 0$.

Local polynomial regression extends kernel estimation to a polynomial fit at the focal point x_0, using local kernel weights, $w_i = K[(x_i - x_0)/h]$. The resulting weighted least-squares (WLS) regression fits the equation

$$y_i = a + b_1(x_i - x_0) + b_2(x_i - x_0)^2 + \cdots + b_p(x_i - x_0)^p + e_i$$

to minimize the weighted residual sum of squares, $\sum_{i=1}^{n} w_i e_i^2$.

- Once the WLS solution is obtained, the fitted value at the focal x_0 is just $\widehat{y}|x_0 = a$.
- As in kernel estimation, this procedure is repeated for representative focal values of x, or at the observations x_i.
- The bandwidth h can either be fixed or it can vary as a function of the focal x.
- When the bandwidth defines a window of nearest neighbors, as is the case for tricube weights, it is convenient to specify the degree of smoothing by the proportion of observations included in the window. This fraction s is called the *span* of the local-regression smoother.

- The number of observations included in each window is then $m = [sn]$, where the square brackets denote rounding to the nearest whole number.

Selecting $p = 1$ produces a local linear fit, the most common case.

- The 'tilt' of the local linear fit promises reduced bias in comparison with the kernel estimator, which corresponds to $p = 0$. This advantage is most apparent at the boundaries, where the kernel estimator tends to flatten.
- The values $p = 2$ or $p = 3$, local quadratic or cubic fits, produce more flexible regressions. Greater flexibility has the potential to reduce bias further, but flexibility also entails the cost of greater variation.
- There is a theoretical advantage to odd-order local polynomials, so $p = 1$ is generally preferred to $p = 0$, and $p = 3$ to $p = 2$.

Figure 9 illustrates the computation of a local linear regression fit to the Canadian occupational prestige data, using the tricube kernel function and nearest-neighbour bandwidths.

- Panel (a) shows a window corresponding to a span of $.4$, accommodating the $[.4 \times 102] = 40$ nearest neighbors of the focal value $x_{(80)}$.
- Panel (b) shows the tricube weight function defined on this window.
- The locally weighted linear fit appears in panel (c).
- Fitted values calculated at each observed x are connected in panel (d). There is no flattening of the fitted regression function, as there was for kernel estimation.

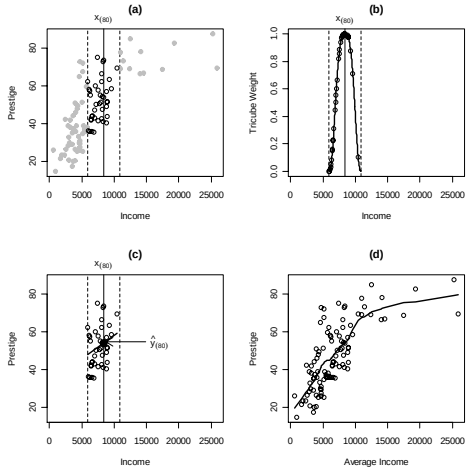

Figure 9. Nearest-neighbor local linear regression of prestige on income.

3.3.1 Selecting the Span by Visual Trial and Error

I will assume nearest-neighbour bandwidths, so bandwidth choice is equivalent to selecting the span of the local-regression smoother. For simplicity, I will also assume a locally linear fit.

A generally effective approach to selecting the span is guided trial and error.

- The span $s = .5$ is often a good point of departure.
- If the fitted regression looks too rough, then try increasing the span; if it looks smooth, then see if the span can be decreased without making the fit too rough.
- We want the *smallest* value of s that provides a smooth fit.

An illustration, for the Canadian occupational prestige data, appears in Figure 10. For these data, selecting $s = .5$ or $s = .7$ appears to provide a reasonable compromise between smoothness and fidelity to the data.

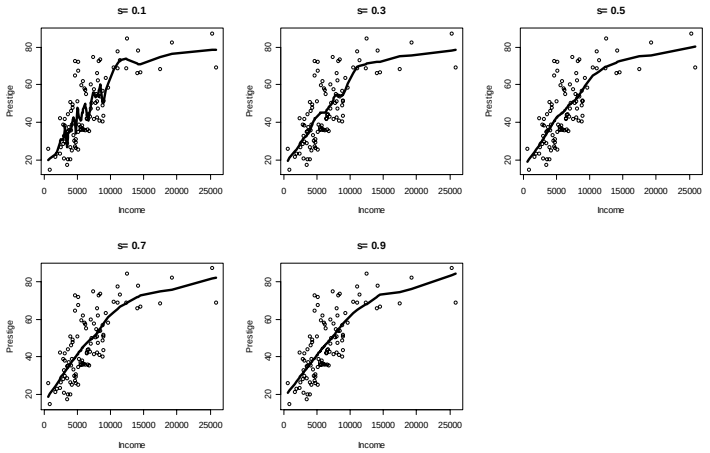

Figure 10. Nearest-neighbor local linear regression of prestige on income, for several values of the span s.

3.3.2 Selecting the Span by Cross-Validation*

A conceptually appealing, but complex, approach to bandwidth selection is to estimate the optimal h (say h^*). We either need to estimate $h^*(x_0)$ for each value x_0 of x at which $\widehat{y}|x$ is to be evaluated, or to estimate an optimal average value to be used with the fixed-bandwidth estimator. A similar approach is applicable to the nearest-neighbour local-regression estimator.

- The so-called *plug-in estimate* of h^* proceeds by estimating its components, which are the error variance σ^2, the curvature of the regression function at the focal x_0, and the density of x-values at x_0. To do this requires a preliminary estimate of the regression function.

- A simpler approach is to estimate the optimal bandwidth or span by *cross-validation* In cross-validation, we evaluate the regression function at the observations x_i.
  - The key idea in cross-validation is to *omit* the ith observation from the local regression at the focal value x_i. We denote the resulting estimate of $E(y|x_i)$ as $\widehat{y}_{-i}|x_i$. Omitting the ith observation makes the fitted value $\widehat{y}_{-i}|x_i$ independent of the observed value y_i.
  - The *cross-validation function* is
$$\text{CV}(s) = \frac{\sum_{i=1}^{n} [\widehat{y}_{-i}(s) - y_i]^2}{n}$$
where $\widehat{y}_{-i}(s)$ is $\widehat{y}_{-i}|x_i$ for span s. The object is to find the value of s that minimizes CV.
  - In practice, we need to compute CV(s) for a range of values of s.
  - Other than repeating the local-regression fit for different values of s, cross-validation does not increase the burden of computation, because we typically evaluate the local regression at each x_i anyway.

  - Although cross-validation is often a useful method for selecting the span, CV(s) is only an estimate, and is therefore subject to sampling variation. Particularly in small samples, this variability can be substantial. Moreover, the approximations to the expectation and variance of the local-regression estimator are asymptotic, and in small samples CV(s) often provides values of s that are too small.
  - There are sophisticated generalizations of cross-validation that are better behaved.

Figure 11 shows CV(s) for the regression of occupational prestige on income. In this case, the cross-validation function provides little specific help in selecting the span, suggesting simply that s should be relatively large. Compare this with the value $s \simeq .6$ that we arrived at by visual trial and error.

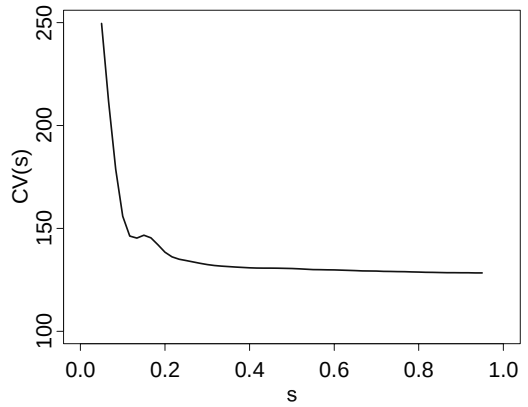

Figure 11. Cross-validation function for the local linear regression of prestige on income.

3.4 Making Local Regression Resistant to Outliers*

As in linear least-squares regression, outliers — and the heavy-tailed error distributions that generate them — can wreak havoc with the local-regression least-squares estimator.

- One solution is to down-weight outlying observations. In linear regression, this strategy leads to M*-estimation*, a kind of robust regression.
- The same strategy is applicable to local polynomial regression.

Suppose that we fit a local regression to the data, obtaining estimates $\widehat{y}_i$ and residuals $e_i = y_i - \widehat{y}_i$.

- Large residuals represent observations that are relatively remote from the fitted regression.

- Now define weights $W_i = W(e_i)$, where the symmetric function $W(\cdot)$ assigns maximum weight to residuals of 0, and decreasing weight as the absolute residuals grow.
  - One popular choice of weight function is the *bisquare* or *biweight*:

$$W_i = W_B(e_i) = \begin{cases} \left[1 - \left(\frac{e_i}{cS}\right)^2\right]^2 & \text{for } |e_i| < cS \\ 0 & \text{for } |e_i| \geq cS \end{cases}$$

  where S is a measure of spread of the residuals, such as $S = \text{median}|e_i|$; and c is a *tuning constant*.
  - Smaller values of c produce greater resistance to outliers but lower efficiency when the errors are normally distributed.
  - Selecting $c = 7$ and using the median absolute deviation produces about 95-percent efficiency compared with least-squares when the errors are normal; the slightly smaller value $c = 6$ is usually used.

  - Another common choice is the *Huber weight function:*

$$W_i = W_H(e_i) = \begin{cases} 1 & \text{for } |e_i| \leq cS \\ cS/|e_i| & \text{for } |e_i| > cS \end{cases}$$

  Unlike the biweight, the Huber weight function never quite reaches 0.
  - The tuning constant $c = 2$ produces roughly 95-percent efficiency for normally distributed errors.

The bisquare and Huber weight functions are graphed in Figure 12.

- We refit the local regression at the focal values x_i by WLS, minimizing the weighted residual sum of squares $\sum_{i=1}^{n} w_i W_i e_i^2$, where the W_i are the 'robustness' weights, just defined, and the w_i are the kernel 'neighborhood' weights.
- Because an outlier will influence the initial local fits, residuals and robustness weights, it is necessary to iterate this procedure until the fitted values $\widehat{y}_i$ stop changing. Two to four robustness iterations almost always suffice.

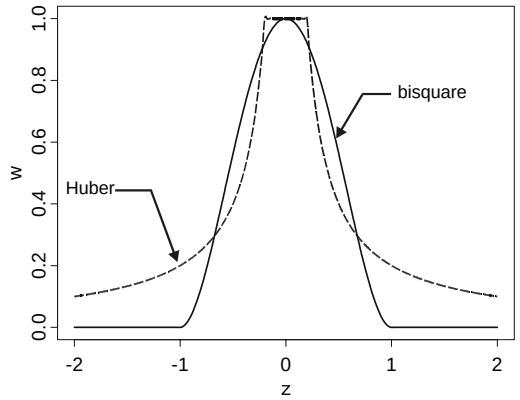

Figure 12. The bisquare (solid line) and Huber (broken line, rescaled) weight functions.

Recall the United Nations data on infant mortality and GDP per capita for 193 countries. Figure 13 shows robust and non-robust local linear regressions of log infant mortality on log GDP. The non-robust fit is pulled towards relatively extreme observations such as Tonga.

Local regression with nearest-neighbour tricube weights and bisquare robustness weights was introduced by Cleveland (1979), who called the procedure *lowess*, for *lo*cally *we*ighted *s*catterplot *s*moothing.

- Upon generalizing the method to multiple regression, Cleveland, Grosse, and Shyu (1992) rechristened it *loess*, for *lo*cal regr*ess*ion.
- Lowess (or loess) is the most widely available method of nonparametric regression.

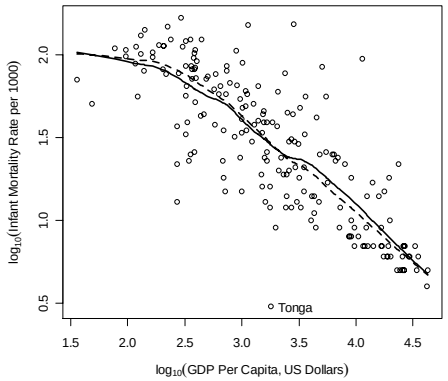

Figure 13. Non-robust (solid line) and robust (broken line) local linear regressions of log infant-mortality rates on log GDP per capita. Both fits use the span $s = .4$.

3.5 Statistical Inference for Local Polynomial Regression

In parametric regression, the central objects of estimation are the regression coefficients. Statistical inference naturally focuses on these coefficients, typically taking the form of confidence intervals or hypothesis tests.

- In nonparametric regression, there are no regression coefficients. The central object of estimation is the regression function, and inference focuses on the regression function directly.
- Many applications of nonparametric regression with one predictor simply have as their goal visual smoothing of a scatterplot. In these instances, statistical inference is at best of secondary interest.

3.5.1 Confidence Envelopes*

Consider the local polynomial estimate $\widehat{y}|x$ of the regression function $f(x)$. For notational convenience, I assume that the regression function is evaluated at the observed predictor values, $x_1, x_2, ..., x_n$.

- The fitted value $\widehat{y}_i = \widehat{y}|x_i$ results from a locally weighted least-squares regression of y on the x values. This fitted value is therefore a weighted sum of the observations:

$$\widehat{y}_i = \sum_{j=1}^{n} s_{ij} y_j$$

where the weights s_{ij} are functions of the x-values.

- Because (by assumption) the y_i's are independently distributed, with common conditional variance $V(y|x = x_i) = V(y_i) = \sigma^2$, the sampling variance of the fitted value $\widehat{y}_i$ is

$$V(\widehat{y}_i) = \sigma^2 \sum_{j=1}^{n} s_{ij}^2$$

- To apply this result, we require an estimate of σ^2. In linear least-squares simple regression, we estimate the error variance as

$$S^2 = \frac{\sum e_i^2}{n-2}$$

where $e_i = y_i - \widehat{y}_i$ is the residual for observation i, and $n - 2$ is the degrees of freedom associated with the residual sum of squares.

- We can calculate residuals in nonparametric regression in the same manner — that is, $e_i = y_i - \widehat{y}_i$.

- To complete the analogy, we require the *equivalent number of parameters* or *equivalent degrees of freedom* for the model, df_{mod}, from which we can obtain the residual degrees of freedom, $df_{\text{res}} = n - df_{\text{mod}}$.
- Then, the estimated error variance is

$$S^2 = \frac{\sum e_i^2}{df_{\text{res}}}$$

and the estimated variance of the fitted value $\widehat{y}_i$ at $x = x_i$ is

$$\widehat{V}(\widehat{y}_i) = S^2 \sum_{j=1}^{n} s_{ij}^2$$

- Assuming normally distributed errors, or a sufficiently large sample, a 95-percent confidence interval for $E(y|x_i) = f(x_i)$ is approximately $\widehat{y}_i \pm 2\sqrt{\widehat{V}(\widehat{y}_i)}$

- Putting the confidence intervals together for $x = x_1, x_2, ..., x_n$ produces a *pointwise 95-percent confidence band* or *confidence envelope* for the regression function.

An example, employing the local linear regression of prestige on income in the Canadian occupational prestige data (with span $s = .6$), appears in Figure 14. Here, $df_{\text{mod}} = 5.0$, and $S^2 = 12,004.72/(102 - 5.0) = 123.76$. The nonparametric-regression smooth therefore uses the equivalent of 5 parameters.

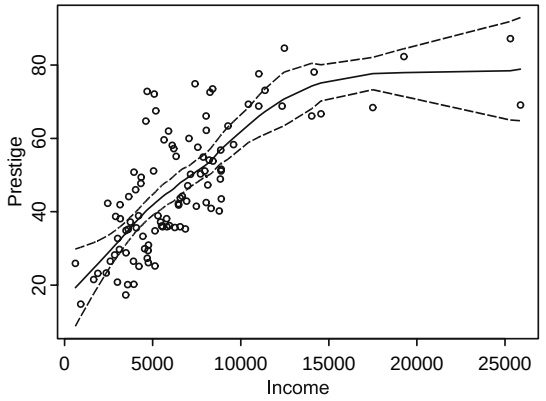

Figure 14. Local linear regression of occupational prestige on income, showing an approximate point-wise 95-percent confidence envelope.

The following three points should be noted:

1. Although the locally linear fit uses the *equivalent* of 5 parameters, it does not produce the same regression curve as fitting a global fourth-degree polynomial to the data.
2. In this instance, the equivalent number of parameters rounds to an integer, but this is an accident of the example.
3. Because $\widehat{y}|x$ is a biased estimate of $E(y|x)$, it is more accurate to describe the envelope around the sample regression as a *"variability band"* rather than as a confidence band.

3.5.2 Hypothesis Tests

In linear least-squares regression, F-tests of hypotheses are formulated by comparing alternative nested models.

- To say that two models are nested means that one, the more specific model, is a special case of the other, more general model.
- For example, in least-squares linear simple regression, the F-statistic

$$F = \frac{\text{TSS} - \text{RSS}}{\text{RSS}/(n-2)}$$

with 1 and $n-2$ degrees of freedom tests the hypothesis of no linear relationship between y and x.

  - The total sum of squares, TSS $= \sum(y_i - \overline{y})^2$, is the variation in y associated with the null model of no relationship, $y_i = \alpha + \varepsilon_i$;
  - the residual sum of squares, RSS $= \sum(y_i - \widehat{y}_i)^2$, represents the variation in y conditional on the linear relationship between y and x, based the model $y_i = \alpha + \beta x_i + \varepsilon_i$.

  - Because the null model is a special case of the linear model, with $\beta = 0$, the two models are (heuristically) nested.

- An analogous, but more general, F-test of no relationship for the nonparametric-regression model is

$$F = \frac{(\text{TSS} - \text{RSS})/(df_{\text{mod}} - 1)}{\text{RSS}/df_{\text{res}}}$$

with $df_{\text{mod}} - 1$ and $df_{\text{res}} = n - df_{\text{mod}}$ degrees of freedom.

  - Here RSS is the residual sum of squares for the nonparametric regression model.
  - Applied to the local linear regression of prestige on income, using the `loess` function in R with a span of 0.6, where $n = 102$, TSS $= 29,895.43$, RSS $= 12,041.37$, and $df_{\text{mod}} = 4.3$, we have

$$F = \frac{(29,895.43 - 12,041.37)/(4.3-1)}{12,041.37/(102-4.3)} = 43.90$$

with $4.3 - 1 = 3.3$ and $102 - 4.3 = 97.7$ degrees of freedom. The resulting p-value is much smaller than $.0001$.

- A test of nonlinearity is simply constructed by contrasting the nonparametric-regression model with the linear simple-regression model.
  - The models are properly nested because a linear relationship is a special case of a general, potentially nonlinear, relationship.
  - Denoting the residual sum of squares from the linear model as RSS_0 and the residual sum of squares from the nonparametric regression model as RSS_1,
$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(df_{\text{mod}} - 2)}{\text{RSS}_1/df_{\text{res}}}$$
with $df_{\text{mod}} - 2$ and $df_{\text{res}} = n - df_{\text{mod}}$ degrees of freedom.
  - This test is constructed according to the rule that the most general model — here the nonparametric-regression model — is employed for estimating the error variance, $S^2 = \text{RSS}_1/df_{\text{res}}$.

  - For the regression of occupational prestige on income, $\text{RSS}_0 = 14,616.17$, $\text{RSS}_1 = 12,004.72$, and $df_{\text{mod}} = 5.0$; thus
$$F = \frac{(14,616.17 - 12,041.37)/(4.3 - 2)}{12,041.37/(102 - 4.3)} = 9.08$$
with $4.3 - 2 = 2.3$ and $102 - 4.3 = 97.7$ degrees of freedom. The corresponding p-value, approximately $.0001$, suggests that the relationship between the two variables is significantly nonlinear.

4. Splines*

Splines are piecewise polynomial functions that are constrained to join smoothly at points called *knots*.

- The traditional use of splines is for interpolation, but they can also be employed for parametric and nonparametric regression.
- Most applications employ cubic splines, the case that I will consider here.
- In addition to providing an alternative to local polynomial regression, smoothing splines are attractive as components of additive regression models and generalized additive models.

4.1 Regression Splines

One approach to simple-regression modeling is to fit a relatively high-degree polynomial in x,
$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$
capable of capturing relationships of widely varying form.

- General polynomial fits, however, are highly nonlocal: Data in one region can substantially affect the fit far away from that region.
- As well, estimates of high-degree polynomials are subject to considerable sampling variation.
- An illustration, employing a cubic polynomial for the regression of occupational prestige on income:
  - Here, the cubic fit does quite well (but dips slightly at the right).

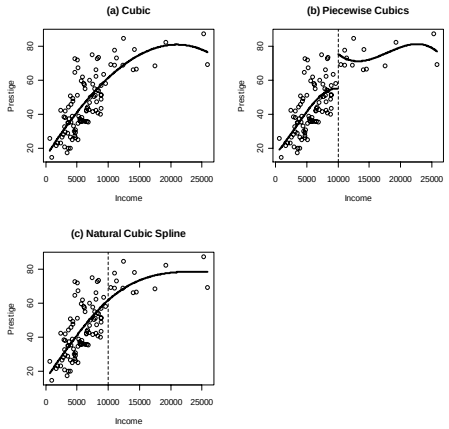

Figure 15. Polynomial fits to the Canadian occupational prestige data: (a) a global cubic fit; (b) independent cubic fits in two bins, divided at Income $= 10,000$; (c) a natural cubic spline, with one knot at Income $= 10,000$.

As an alternative, we can partition the data into bins, fitting a different polynomial regression in each bin.

- A defect of this *piecewise* procedure is that the curves fit to the different bins will almost surely be discontinuous, as illustrated in Figure 15 (b).
- *Cubic regression splines* fit a third-degree polynomial in each bin under the added constraints that the curves join at the bin boundaries (the knots), and that the first and second derivatives (i.e., the slope and curvature of the regression function) are continuous at the knots.
- *Natural cubic regression splines* add knots at the boundaries of the data, and impose the additional constraint that the fit is linear beyond the terminal knots.
  - This requirement tends to avoid wild behavior near the extremes of the data.
  - If there are k 'interior' knots and two knots at the boundaries, the natural spline uses $k + 2$ independent parameters.

  - With the values of the knots fixed, a regression spline is just a linear model, and as such provides a fully parametric fit to the data.
  - Figure 15 (c) shows the result of fitting a natural cubic regression spline with one knot at Income $= 10,000$, the location of which was determined by examining the scatterplot, and the model therefore uses only 3 parameters.

4.2 Smoothing Splines

In contrast to regression splines, *smoothing splines* arise as the solution to the following nonparametric-regression problem: Find the function $\widehat{f}(x)$ with two continuous derivatives that minimizes the *penalized sum of squares*,

$$\mathsf{SS}^*(h) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + h \int_{x_{\min}}^{x_{\max}} [f''(x)]^2 \, dx$$

where h is a smoothing constant, analogous to the bandwidth of a kernel or local-polynomial estimator.

- The first term in the equation is the residual sum of squares.
- The second term is a *roughness penalty,* which is large when the integrated second derivative of the regression function $f''(x)$ is large — that is, when $f(x)$ is rough.

- If $h = 0$ then $\widehat{f}(x)$ simply interpolates the data.
- If h is very large, then $\widehat{f}$ will be selected so that $\widehat{f}''(x)$ is everywhere 0, which implies a globally linear least-squares fit to the data.

It turns out that the function $\widehat{f}(x)$ that minimizes $SS^*(h)$ is a natural cubic spline with knots at the distinct observed values of x.

- Although this result seems to imply that n parameters are required, the roughness penalty imposes additional constraints on the solution, typically reducing the equivalent number of parameters for the smoothing spline greatly.
  - It is common to select the smoothing constant h indirectly by setting the equivalent number of parameters for the smoother.
  - An illustration appears in Figure 16, comparing a smoothing spline with a local-linear fit employing the same equivalent number of parameters (degrees of freedom).
- Smoothing splines offer certain small advantages in comparison with local polynomial smoothers, but generally provide similar results.

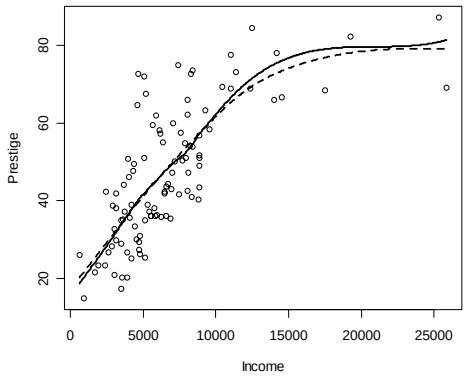

Figure 16. Nonparametric regression of occupational prestige on income, using local linear regression (solid line) and a smoothing spline (broken line), both with 4.3 equivalent parameters.

5. Nonparametric Regression and Data Analysis*

The scatterplot is the most important data-analytic statistical graph. I am tempted to suggest that you add a nonparametric-regression smooth to every scatterplot that you draw, since the smooth will help to reveal the relationship between the two variables in the plot.

Because scatterplots are adaptable to so many different contexts in data analysis, it is not possible to exhaustively survey their uses here. Instead, I will concentrate on an issue closely related to nonparametric regression: Detecting and dealing with nonlinearity in regression analysis.

- One response to the possibility of nonlinearity is to employ nonparametric multiple regression.
- An alternative is to fit a preliminary linear regression; to employ appropriate diagnostic plots to detect departures from linearity; and to follow up by specifying a new parametric model that captures nonlinearity detected in the diagnostics, for example by transforming a predictor.

5.1 The 'Bulging Rule'

My first example examined the relationship between the infant-mortality rates and GDP per capita of 193 nations of the world.

- A scatterplot of the data supplemented by a local-linear smooth, in Figure 1 (a), reveals a highly nonlinear relationship between the two variables: Infant mortality declines smoothly with GDP, but at a rapidly decreasing rate.
- Taking the logarithms of the two variables, in Figure 17 (b), renders the relationship nearly linear.

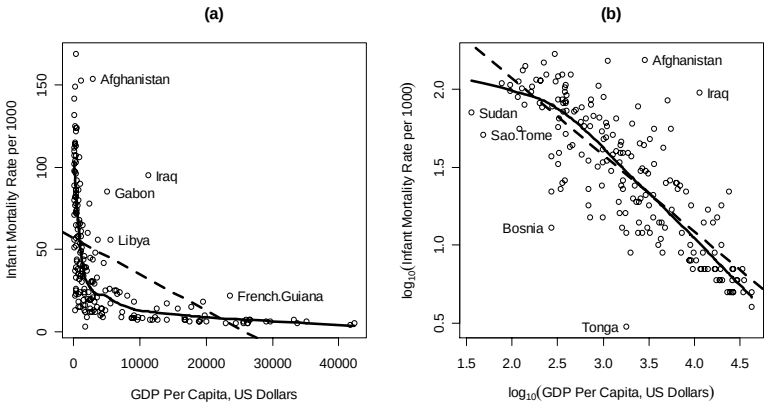

Figure 17. Infant-mortality rate per 1000 and GDP per capita (US dollars) for 193 nations. (Figure 1 repeated.)

Mosteller and Tukey (1977) suggest a systematic rule — which they call the *'bulging rule'* — for selecting linearizing transformations from the *family of powers and roots*, where a variable x is replaced by the power x^p.

- For example, when $p = 2$, the variable is replaced by its square, x^2; when $p = -1$, the variable is replaced by its inverse, $x^{-1} = 1/x$; when $p = 1/2$, the variable is replaced by its square-root, $x^{1/2} = \sqrt{x}$; and so on.
- The only exception to this straightforward definition is that $p = 0$ designates the log transformation, $\log x$, rather than the 0th power.
- We are not constrained to pick simple values of p, but doing so often aids interpretation.

- Transformations in the family of powers and roots are only applicable when all of the values of x are positive:
  - Some of the transformations, such as square-root and log, are undefined for negative values of x.
  - Other transformations, such as x^2, would distort the order of x if some x-values are negative and some are positive.
- A simple solution is to use a 'start' — to add a constant quantity c to all values of x prior to applying the power transformation: $x \longrightarrow (x + c)^p$.
- Notice that negative powers — such as the inverse transformation, x^{-1} — reverse the order of the x-values; if we want to preserve the original order, then we can take $x \longrightarrow -x^p$ when p is negative.
- Alternatively, we can use the similarly shaped *Box-Cox family of transformations*:

$$x \longrightarrow x^{(p)} = \begin{cases} (x^p - 1)/p & \text{for } p \neq 0 \\ \log_e x & \text{for } p = 0 \end{cases}$$

Power transformation of x or y can help linearize a nonlinear relationship that is both *simple* and *monotone*. What is meant by these terms is illustrated in Figure 18:

- A relationship is simple when it is smoothly curved and when the curvature does not change direction.
- A relationship is monotone when y strictly increases or decreases with x.
  - Thus, the relationship in Figure 18 (a) is simple and monotone;
  - the relationship in Figure 18 (b) is monotone but not simple, since the direction of curvature changes from opening up to opening down;
  - the relationship in Figure 18 (c) is simple but *non*monotone, since y first decreases and then increases with x.

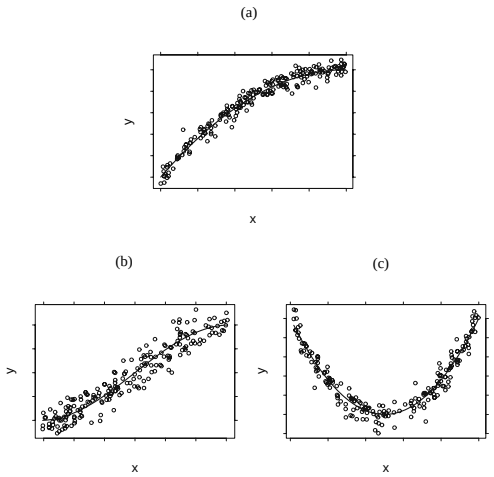

Figure 18. The relationship in (a) is simple and monotone; that in (b) is montone but not simple; and that in (c) is simple but nonmonotone.

Although nonlinear relationships that are not simple or that are nonmonotone cannot be linearized by a power transformation, other forms of parametric regression may be applicable. For example, the relationship in Figure 18 (c) could be modeled as a quadratic equation:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- Polynomial regression models, such as quadratic equations, can be fit by linear least-squares regression.
- Nonlinear least squares can be used to fit an even broader class of parametric models.

Mosteller and Tukey's bulging rule is illustrated in Figure 19:

- When, as in the infant-mortality data of Figure 1 (a), the bulge points *down* and to the *left*, the relationship is linearized by moving x '*down* the ladder' of powers and roots, towards $\sqrt{x}$, $\log x$, and $1/x$, or moving y *down* the ladder of powers and roots, or both.
- When the bulge points *up*, we can move x *up* the ladder of powers, towards x^2 and x^3.
- When the bulge points to the *right*, we can move y *up* the ladder of powers.
- Specific linearizing transformations are located by trial and error; the farther one moves from no transformation ($p = 1$), the greater the effect of the transformation.

In the example, log transformations of both infant mortality and GDP somewhat overcorrect the original nonlinearity, producing a small bulge pointing up and to the right.

- Nevertheless, the nonlinearity in the transformed data is relatively slight, and using log transformations for both variables yields a simple interpretation.
- The straight line plotted in Figure 17 (b) has the equation

$$\log_{10}\widehat{\text{Infant Mortality}} = 3.06 - 0.493 \times \log_{10}\text{GDP}$$

- The slope of this relationship, $b = -0.493$, is what economists call an *elasticity*: On average, a one-percent increase in GDP per capita is associated with an approximate one-half-percent decline in the infant-mortality rate.

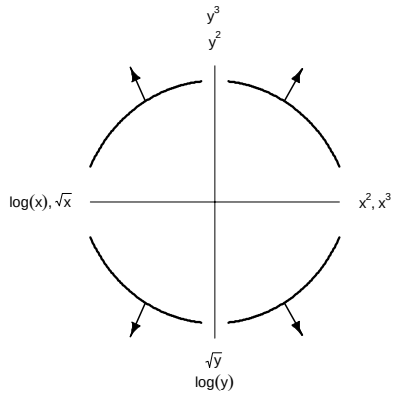

Figure 19. Mosteller and Tukey's 'bulging rule' for locating a linearizing transformation.

5.2 Component+Residual Plots

Suppose that y is additively, but not necessarily linearly, related to $x_1, x_2, ..., x_k$, so that

$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki}) + \varepsilon_i$$

- If the partial-regression function f_j is simple and monotone, then we can use the bulging rule to find a transformation that linearizes the partial relationship between y and the predictor x_j.
- Alternatively, if f_j takes the form of a simple polynomial in x_j, such as a quadratic or cubic, then we can specify a parametric model containing polynomial terms in that predictor.

Discovering nonlinearity in multiple regression is more difficult than in simple regression because the predictors typically are correlated. The scatterplot of y against x_j is informative about the *marginal* relationship between these variables, ignoring the other predictors, not necessarily about the *partial* relationship f_j of y to x_j, holding the other x's constant.

Under relatively broad circumstances *component+residual plots* (also called *partial-residual plots*) can help to detect nonlinearity in multiple regression.

- We fit a preliminary linear least-squares regression,

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i$$

- The *partial residuals* for x_j add the least-squares residuals to the linear component of the relationship between y and x_j:

$$e_{i[j]} = e_i + b_j x_{ji}$$

- An unmodeled nonlinear component of the relationship between y and x_j should appear in the least-squares residuals, so plotting and smoothing $e_{[j]}$ against x_j will reveal the partial relationship between y and x_j. We think of the smoothed partial-residual plot as an estimate $\widehat{f}_j$ of the partial-regression function.
- This procedure is repeated for each predictor, $j = 1, 2, ..., k$.

Illustrative component+residual plots appear in Figure 20, for the regression of prestige on income and education.

- The solid line on each plot gives a local-linear fit for span $s = .6$.
- the broken line gives the linear least-squares fit, and represents the least-squares multiple-regression plane viewed edge-on in the direction of the corresponding predictor.

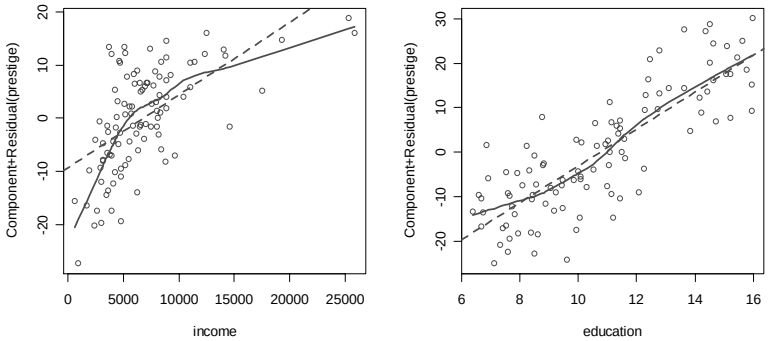

Figure 20. Component+residual plots for the regression of occupational prestige on income and education.

- The left panel shows that the partial relationship between prestige and income controlling for education is substantially nonlinear. Although the nonparametric regression curve fit to the plot is not altogether smooth, the bulge points up and to the left, suggesting transforming income down the ladder of powers and roots. Visual trial and error indicates that the log transformation of income serves to straighten the relationship between prestige and income.
- The right panel suggests that the partial relationship between prestige and education is nonlinear and monotone, but not simple. Consequently, a power transformation of education is not promising. We could try specifying a cubic regression for education (including education, education^2, and education^3 in the regression model), but the departure from linearity is slight, and a viable alternative here is simply to treat the education effect as linear.

- Regressing occupational prestige on education and the log (base 2) of income produces the following result:

$$\widehat{\text{Prestige}} = -95.2 + 7.93 \times \log_2 \text{Income} + 4.00 \times \text{Education}$$

  – Holding education constant, doubling income (i.e., increasing $\log_2$Income by 1) is associated on average with an increment in prestige of about 8 points;

  – holding income constant, increasing education by 1 year is associated on average with an increment in prestige of 4 points.

6. Nonparametric Multiple Regression

I will describe two generalizations of nonparametric regression to two or more predictors:

1. The local polynomial multiple-regression smoother, which fits the general model

$$y_i = f(x_{1i}, x_{2i}, ..., x_{ki}) + \varepsilon_i$$

2. The additive nonparametric regression model

$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki}) + \varepsilon_i$$

6.1 Local Polynomial Multiple Regression

As a formal matter, it is simple to extend the local-polynomial estimator to several predictors:

- To obtain a fitted value $\widehat{y}|\mathbf{x}_0$ at the focal point $\mathbf{x}_0 = (x_{1,0}, x_{2,0}, ..., x_{k0})'$ in the predictor space, we perform a weighted-least-squares polynomial regression of y on the x's, emphasizing observations close to the focal point.
  – A local linear fit takes the form:

$$\begin{aligned} y_i &= a + b_1(x_{1i} - x_{1,0}) + b_2(x_{2i} - x_{2,0}) \\ &\quad + \cdots + b_k(x_{ki} - x_{k0}) + e_i \end{aligned}$$

  – For $k = 2$ predictors, a local quadratic fit takes the form

$$\begin{aligned} y_i &= a + b_1(x_{1i} - x_{1,0}) + b_2(x_{2i} - x_{2,0}) \\ &\quad + b_{11}(x_{1i} - x_{1,0})^2 + b_{22}(x_{2i} - x_{2,0})^2 \\ &\quad + b_{12}(x_{1i} - x_{1,0})(x_{2i} - x_{2,0}) + e_i \end{aligned}$$

  When there are several predictors, the number of terms in the local quadratic regression grows large, and consequently I will not consider cubic or higher-order polynomials.

  – In either the linear or quadratic case, we minimize the weighted sum of squares $\sum_{i=1}^{n} w_i e_i^2$ for suitably defined weights w_i. The fitted value at the focal point in the predictor space is then $\widehat{y}|\mathbf{x}_0 = a$.

6.1.1 Finding Kernel Weights in Multiple Regression*

- There are two straightforward ways to extend kernel weighting to local polynomial multiple regression:

  (a) Calculate *marginal weights* separately for each predictor,

  $$w_{ij} = K[(x_{ji} - x_{j0})/h_j]$$

  Then

  $$w_i = w_{i1}w_{i2}\cdots w_{ik}$$

  (b) Measure the distance $D(\mathbf{x}_i, \mathbf{x}_0)$ in the predictor space between the predictor values $\mathbf{x}_i$ for observation i and the focal $\mathbf{x}_0$. Then

  $$w_i = K\left[\frac{D(\mathbf{x}_i, \mathbf{x}_0)}{h}\right]$$

  There is, however, more than one way to define distances between points in the predictor space:

- *Simple Euclidean distance*:

  $$D_E(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^{k}(x_{ji} - x_{j0})^2}$$

  Euclidean distances only make sense when the x's are measured in the same units (e.g., for spatially distributed data, where the two predictors x_1 and x_2 represents coordinates on a map).

- *Scaled Euclidean distance*: Scaled distances adjust each x by a measure of dispersion to make values of the predictors comparable. For example,

  $$z_{ji} = \frac{x_{ji} - \overline{x}_j}{s_j}$$

  where $\overline{x}_j$ and s_j are the mean and standard deviation of x_j. Then

  $$D_S(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^{k}(z_{ji} - z_{j0})^2}$$

  This is the most common approach to defining distances.

(c) *Generalized distance*: Generalized distances adjust not only for the dispersion of the x's but also for their correlational structure:

$$D_G(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{(\mathbf{x}_i - \mathbf{x}_0)'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{x}_0)}$$

where $\mathbf{V}$ is the covariance matrix of the x's, perhaps estimated robustly. Figure 21 illustrates generalized distances for $k = 2$ predictors.

- As mentioned, simple Euclidean distances do not make sense unless the predictors are on the same scale. Beyond that point, the choice of product marginal weights, weights based on scaled Euclidean distances, or weights based on generalized distances usually does not make a great deal of difference.
- Methods of bandwidth selection and statistical inference for local polynomial multiple regression are essentially identical to the methods discussed previously for nonparametric simple regression.

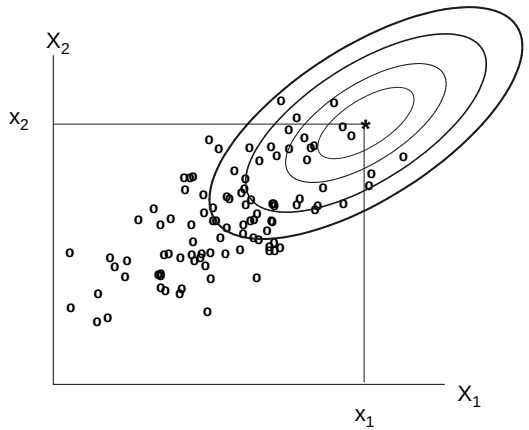

Figure 21. Contours of constant generalized distance from the focal point $\mathbf{x}_0 = (x_1, x_2)'$, represented by the asterisk. Notice that the contours are elliptical.

6.1.2 Obstacles to Nonparametric Multiple Regression

Although it is therefore simple to extend local polynomial estimation to multiple regression, there are two flies in the ointment:

1. *The 'curse of dimensionality'*: As the number of predictors increases, the number of observations in the local neighborhood of a focal point tends to decline rapidly. To include a fixed number of observations in the local fits therefore requires making neighborhoods less and less local.
   – The problem is illustrated in Figure 22 for $k = 2$ predictors. This figure represents a "best-case" scenario, where the x's are independent and uniformly distributed. Neighborhoods constructed by product-marginal weighting correspond to square (more generally, rectangular) regions in the graph. Neighborhoods defined by distance from a focal point correspond to circular (more generally, elliptical) regions in the graph.

   – To include half the observations in a square neighborhood centered on a focal x, we need to define marginal neighborhoods for each of x_1 and x_2 that include roughly $\sqrt{1/2} \simeq .71$ of the data; for $k = 10$ predictors, the marginal neighborhoods corresponding to a hyper-cube that encloses half the observations would each include about $\sqrt[10]{1/2} \simeq 0.93$ of the data.
   – A circular neighborhood in two dimensions enclosing half the data has diameter $2\sqrt{0.5/\pi} \simeq 0.8$ along each axis; the diameter of the hyper-sphere enclosing half the data also grows with dimensionality, but the formula is complicated.

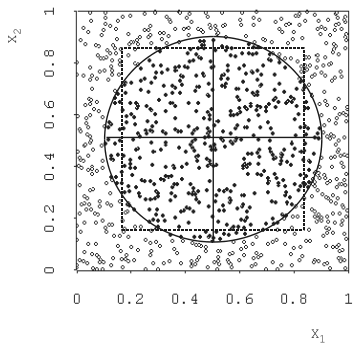

Figure 22. The 'curse of dimensionality': 1,000 observations for independent, uniformly distributed random variables x_1 and x_2. The 500 nearest neighbors of the focal point $\mathbf{x}_0 = (.5, .5)'$ are highlighted, along with the circle that encloses them. Also shown is the square centered on $\mathbf{x}_0$ enclosing half the data.

2. *Difficulties of interpretation*: Because nonparametric regression does not provide an equation relating the average response to the predictors, we need to display the response surface graphically.
   – This is no problem when there is only one x, since the scatterplot relating y to x is two-dimensional and the regression "surface" is just a curve.
   – When there are two x's, the scatterplot is three-dimensional and the regression surface is two-dimensional. Here, we can represent the regression surface in an isometric or perspective plot, as a contour plot, or by slicing the surface. These strategies are illustrated in an example below.
   – Although slicing can be extended to more predictors, the result becomes difficult to examine, particularly when the number of predictors exceeds three.
   – These problems motivate the additive regression model (to be described later).

6.1.3 An Example: The Canadian Occupational Prestige Data

To illustrate local polynomial multiple regression, let us return to the Canadian occupational prestige data, regressing prestige on the income and education levels of the occupations.

- Local quadratic and local linear fits to the data using the `loess` function in R produce the following numbers of equivalent parameters (df_{mod}) and residual sums of squares:

Model	df_{mod}	*RSS*
Local linear	8.0	4245.9
Local quadratic	15.4	4061.8

The span of the local-polynomial smoothers, $s = .5$ (corresponding roughly to marginal spans of $\sqrt{.5} \simeq .7$), was selected by visual trial and error.

- An incremental F-test for the extra terms in the quadratic fit is

$$F = \frac{(4245.9 - 4061.8)/(15.4 - 8.0)}{4061.8/(102 - 15.4)} = 0.40$$

with $15.4 - 8.0 = 7.4$ and $102 - 15.4 = 86.6$ degrees of freedom, for which $p = .89$, suggesting that little is gained from the quadratic fit.

- Figures 23–26 show three graphical representations of the local linear fit:

(a) Figure 23 is a *perspective plot* of the fitted regression surface. It is relatively easy to visualize the general relationship of prestige to education and income, but hard to make precise visual judgments:
   - Prestige generally rises with education at fixed levels of income.
   - Prestige rises with income at fixed levels of education, at least until income gets relatively high.
   - But it is difficult to discern, for example, the fitted value of prestige for an occupation at an income level of $10,000 and an education level of 12 years.

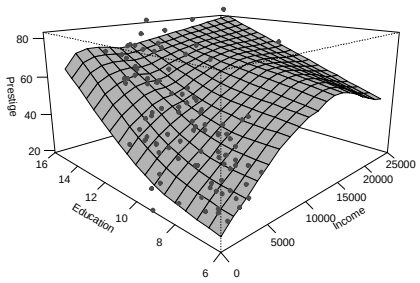

Figure 23. Perspective plot for the local-linear regression of occupational prestige on income and education.

(b) Figure 24 is a *contour plot* of the data, showing "iso-prestige" lines for combinations of values of income and education.

- I find it difficult to visualize the regression surface from a contour plot (perhaps hikers and mountain climbers do better).
- But it is relatively easy to see, for example, that our hypothetical occupation with an average income of $10,000 and an average education level of 12 years has fitted prestige between 50 and 60 points.

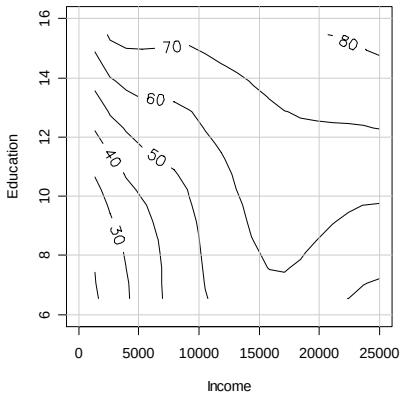

Figure 24. Contour plot for the local linear regression of occupational prestige on income and education.

(c) Figure 25 is a *conditioning plot* or *'coplot'* (due to William Cleveland), showing the fitted relationship between occupational prestige and income for several levels of education.

- The levels at which education is 'held constant' are given in the upper panel of the figure.
- Each of the remaining panels — proceeding from lower left to upper right — shows the fit at a particular level of education.
- These are the lines on the regression surface in the direction of income (fixing education) in the perspective plot (Figure 23), but displayed two-dimensionally.
- The vertical lines give pointwise 95-percent confidence intervals for the fit. The confidence intervals are wide where data are sparse — for example, for occupations at very low levels of education but high levels of income.
- Figure 26 shows a similar coplot displaying the fitted relationship between prestige and education controlling for income.

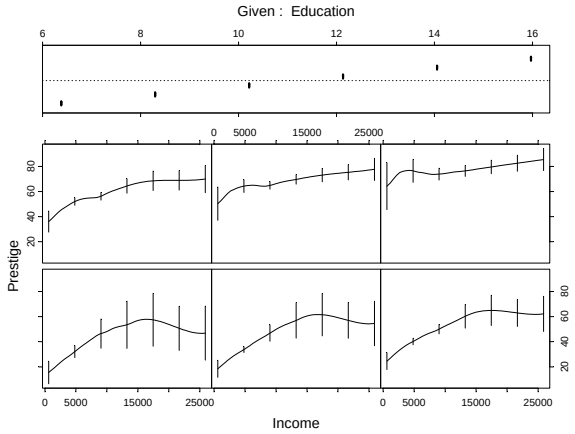

Figure 25. Conditioning plot showing the relationship between occupational prestige and income for various levels of education. (Note: Made with S-PLUS.)

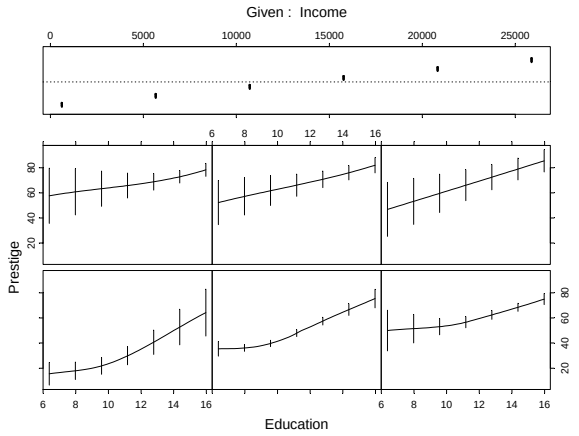

Figure 26. Conditioning plot showing the relationship between occupational prestige and education for various levels of income. (Note: Made with S-PLUS.)

Is prestige significantly related to income and education?

- We can answer this question by dropping each predictor in turn and noting the increase in the residual sum of squares.
- Because the span for the local-linear multiple-regression fit is $s = .5$, the corresponding simple-regression models use spans of $s = \sqrt{.5} \simeq .7$:

Model	df_{mod}	*RSS*
Income and Education	8.0	4245.9
Income	3.8	12,006.1
Education	3.0	7640.2

- F-tests for income and education are as follows:

$$F_{\text{Income}} = \frac{(7640.2 - 4245.9)/(8.0 - 3.0)}{4245.9/(102 - 8.0)} = 7.8$$

$$F_{\text{Education}} = \frac{(12,006.1 - 4245.9)/(8.0 - 3.8)}{4245.9/(102 - 8.0)} = 40.9$$

These F-statistics have, respectively, 5.0 and 94.0 degrees of freedom, and 4.2 and 94.0 degrees of freedom. Both p-values are close to 0.

Perspective plots and contour plots cannot easily be generalized to more than two predictors:

- Although three-dimensional contour plots can be constructed, they are very difficult to examine, in my opinion, and higher-dimensional contour plots are out of the question.
- One can construct two-dimensional perspective or contour plots at fixed combinations of values of other predictors, but the resulting displays are confusing.
- Coplots can be constructed for three predictors by arranging combinations of values of two of the predictors in a rectangular array, and displaying the partial relationship between the response and the third predictor for each such combination. By rotating the role of the third predictor, three coplots are produced.
- Coplots can in principle be generalized to any number of predictors, but the resulting proliferation of graphs quickly gets unwieldy.

6.2 Additive Regression Models

In unrestricted nonparametric multiple regression, we model the conditional average value of y as a general, smooth function of several x's,

$$E(y|x_1, x_2, ..., x_k) = f(x_1, x_2, ..., x_k)$$

- In linear regression analysis, in contrast, the average value of the response variable is modeled as a linear function of the predictors,

$$E(y|x_1, x_2, ..., x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- As the linear model, the *additive regression model* specifies that the average value of y is a sum of separate terms for each predictor, but these terms are merely assumed to be smooth functions of the x's:

$$E(y|x_1, x_2, ..., x_k) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k)$$

The additive regression model is more restrictive than the general nonparametric-regression model, but more flexible than the standard linear-regression model.

- A considerable advantage of the additive regression model is that it reduces to a series of two-dimensional partial-regression problems. This is true both in the computational sense and, even more importantly, with respect to interpretation:
  - Because each partial-regression problem is two-dimensional, we can estimate the partial relationship between y and x_j by using a suitable scatterplot smoother, such as local polynomial regression. We need somehow to remove the effects of the other predictors, however — we cannot simply smooth the scatterplot of y on x_j ignoring the other x's. Details are given later.
  - A two-dimensional plot suffices to examine the estimated partial-regression function $\widehat{f_j}$ relating y to x_j holding the other x's constant.

Figure 27 shows the estimated partial-regression functions for the additive regression of occupational prestige on income and education.

- Each partial-regression function was fit by a nearest-neighbour local-linear smoother, using span $s = .7$.

- The points in each graph are partial residuals for the corresponding predictor, removing the effect of the other predictor.
- The broken lines mark off pointwise 95-percent confidence envelopes for the partial fits.

Figure 28 is a three-dimensional perspective plot of the fitted additive-regression surface relating prestige to income and education.

- Slices of this surface in the direction of income (i.e., holding education constant at various values) are all parallel,
- Llikewise slices in the direction of education (holding income constant) are parallel
- This is the essence of the additive model, ruling out interaction between the predictors. Because all of the slices are parallel, we need only view one of them edge-on, as in Figure 27.
- Compare the additive-regression surface with the fit of the unrestricted nonparametric-regression model in Figure 23.

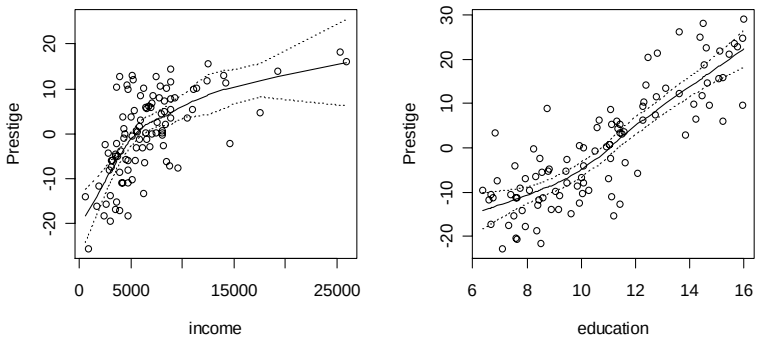

Figure 27. Plots of the estimated partial-regression functions for the additive regression of prestige on income and education.

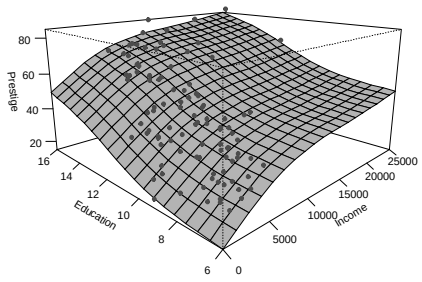

Figure 28. Perspective plot of the fitted additive regression of prestige on income and education.

Is anything lost in moving from the general nonparametric-regression model to the more restrictive additive model?

- Residual sums of squares and equivalent numbers of parameters for the two models are as follows:

Model	df_{mod}	*RSS*
General	8.0	4245.9
Additive	6.9	4658.2

- An approximate F-test comparing the two models is

$$F = \frac{(4658.2 - 4245.9)/(8.0 - 6.9)}{4245.9/(102 - 8.0)} = 8.3$$

with 1.1 and 94.0 degrees of freedom, for which $p = .004$. There is, therefore, evidence of lack of fit for the additive model, although the R^2's for the two models are not very different: .858 and .844, respectively.

Note: The additive model was fit with the `gam` function in the R `gam` package, but the degrees of freedom were corrected to agree with `gam` in S-PLUS.

- To test the contribution of each predictor to the additive model, we compare the full additive model with models omitting each predictor in turn:

Model	df_{mod}	*RSS*
Additive	6.9	4658.2
Income only	4.5	11,981.8
Education only	3.4	7626.9

Then

$$F_{\text{Income}} = \frac{(7626.9 - 4658.2)/(6.9 - 3.4)}{4658.2/(102 - 6.9)} = 17.31$$

$$F_{\text{Education}} = \frac{(11,981.8 - 4658.2)/(6.9 - 4.5)}{4658.2/(102 - 6.9)} = 62.30$$

with, respectively, 3.5 and 95.1, and 2.4 and 95.1 degrees of freedom; both F-statistics have p-values close to 0. Again these results are from the `gam` function in the R `gam` package, but the degrees of freedom are corrected.

6.2.1 Fitting the Additive Model to Data*

For simplicity, consider the case of two predictors:

$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \varepsilon_i$$

- Suppose, unrealistically, that the partial-regression function f_2 is known, but that f_1 is not. Rearranging the regression equation,

$$y_i - f_2(x_{2i}) = \alpha + f_1(x_{1i}) + \varepsilon_i$$

So, smoothing $y_i - f_2(x_{2i})$ against x_{1i} will produce an estimate of $\alpha + f_1(x_{1i})$.

- The regression constant α is a bit of a nuisance. We could absorb α into one of the partial-regression functions, or we could force the partial-regression functions evaluated at the observed x_{ji}'s to sum of 0; in this case, α becomes the unconditional expectation of y, estimated by $\overline{y}$. Then we estimate f_1 by smoothing $y_i - \overline{y} - f_2(x_{2i})$ against x_{1i}.

- In a real application, neither f_1 nor f_2 is known.

(a) Let us start with preliminary estimates, denoted $\widehat{f}_1^{(0)}$ and $\widehat{f}_2^{(0)}$, based on the linear least-squares regression of y on the x's:

$$y_i - \overline{y} = b_1(x_{1i} - \overline{x}_1) + b_2(x_{2i} - \overline{x}_2) + e_i$$

Then

$$\widehat{f}_1^{(0)}(x_{1i}) = b_1(x_{1i} - \overline{x}_1)$$
$$\widehat{f}_2^{(0)}(x_{2i}) = b_2(x_{2i} - \overline{x}_2)$$

Expressing the variables as deviations from their means insures that the partial-regression functions sum to 0.

(b) Form the partial residual

$$\begin{aligned} e^{(1)}_{i[1]} &= y_i - \overline{y} - b_2(x_{2i} - \overline{x}_2) \\ &= e_i + b_1(x_{1i} - \overline{x}_1) \end{aligned}$$

which removes from y its linear relationship to x_2, but retains the linear relationship between y and x_1, possibly along with a nonlinear relationship in the least-squares residuals e_i. Smoothing $e^{(1)}_{i[1]}$ against x_{1i} provides a new estimate $\widehat{f}^{(1)}_1$ of f_1.

(c) Using the estimate $\widehat{f}^{(1)}_1$, form partial residuals for x_2:

$$e^{(1)}_{i[2]} = y_i - \overline{y} - \widehat{f}^{(1)}_1(x_{1i})$$

Smoothing $e^{(1)}_{i[2]}$ against x_{2i} yields a new estimate $\widehat{f}^{(1)}_2$ of f_2.

(d) The new estimate $\widehat{f}^{(1)}_2$, in turn, is used to calculate updated partial residuals $e^{(2)}_{i[1]}$ for x_1, which, when smoothed against x_{1i}, produce the updated estimate $\widehat{f}^{(2)}_1$ of f_1.

(e) This iterative process, called *backfitting*, continues until the estimated partial-regression functions stabilize. In the absence of a generalization of collinearity (which Hastie and Tibshirani term *concurvity*), backfitting is "guaranteed" to converge to a unique solution regardless of the starting partial-regression functions, using either local-regression or spline smoothers.

Backfitting, by the way, is not the only approach to fitting additive regression model.

6.3 Semiparametric Models and Models with Interactions

This section develops two straightforward relatives of additive regression models:

- *Semparametric models* are additive regression models in which some terms enter nonparametrically while others enter linearly.
- Models in which some of the predictors interact, for example in pairwise fashion.

It is, as well, possible to combine these strategies, so that some terms enter linearly, others additively, and still others are permitted to interact.

The semiparametric regression model is written

$$\begin{aligned} y_i = \; & \alpha + \beta_1 x_{1i} + \cdots + \beta_r x_{ri} \\ & + f_{r+1}(x_{r+1,i}) + \cdots + f_k(x_{ki}) + \varepsilon_i \end{aligned}$$

- The first r predictors enter the model linearly, while the partial relationships of y to the remaining $k - r$ predictors are simply assumed to be smooth.
- The semiparametric model can be estimated by backfitting. In each iteration, all of the linear terms can be estimated in a single step: Form partial residuals that remove the current estimates of the nonparametric terms, and then regress these partial residuals on $x_1, ..., x_r$ to obtain updated estimates of the β's.

- The semiparametric model is applicable whenever there is reason to believe that one or more x's enter the regression linearly:
  - In rare instances, there may be prior reasons for believing that this is the case, or examination of the data might suggest a linear relationship, perhaps after transforming an x.
  - More commonly, if some of the x's are dummy variables — representing one or more categorical predictors — then it is natural to enter the dummy variables as linear terms.

- We can test for nonlinearity by contrasting two models, one of which treats a predictor nonparametrically and the other linearly.
  - For example, to test for nonlinearity in the partial relationship between y and x_1, we contrast the additive model
$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki}) + \varepsilon_i$$
with the semiparametric model
$$y_i = \alpha + \beta_1 x_{i1} + f_2(x_{2i}) + \cdots + f_k(x_{ki}) + \varepsilon_i$$

  - To illustrate, let us fit three models for the regression of occupational prestige on income and education:

Model	df_{mod}	*RSS*
1 Additive	6.9	4658.2
2 Income linear	4.4	5675.0
3 Education linear	5.5	4956.0

Model 1 is the additive regression model; model 2 is a semiparametric model containing a linear term for income and a nonparametric term for education; model 3 is a semiparametric model with a linear term for education and a nonparametric term for income. (Once again, the results are from the `gam` function in the R `gam` package, but the degrees of freedom are corrected.)

  - Contrasting models 1 and 2 produces a test for nonlinearity in the partial relationship of prestige to income; contrasting models 1 and 3 produces a test for nonlinearity in the relationship of prestige to education:
$$F_{\text{Income(nonlinear)}} = \frac{(5675.0 - 4658.2)/(6.9 - 4.4)}{4658.2/(102 - 6.9)} = 8.30$$
$$F_{\text{Education(nonlinear)}} = \frac{(4956.0 - 4658.2)/(6.9 - 5.5)}{4658.2/(102 - 6.9)} = 4.34$$
The first of these F-test statistics has 2.5 and 97.1 degrees of freedom, with $p = .0002$; the second has 1.4 and 97.1 degrees of freedom, with $p = .03$.
  - There is, therefore, much stronger evidence of a nonlinear partial relationship between prestige and income than between prestige and education.

While semiparametric regression models make the additive model more restrictive, incorporating interactions makes the model more flexible.

- For example, the following model permits interaction (nonadditivity) in the partial relationship of y to x_1 and x_2:

$$y_i = \alpha + f_{12}(x_{1i}, x_{2i}) + f_3(x_{3i}) + \cdots + f_k(x_{ki}) + \varepsilon_i$$

- Once again, this model can be estimated by backfitting, employing a multiple-regression smoother (such as local polynomial multiple regression) to estimate f_{12}.
- Contrasting this model with the more restrictive additive model produces an incremental F-test for the interaction between x_1 and x_2.
- This strategy can, in principle, be extended to models with higher-order interactions — for example, $f_{123}(x_{1i}, x_{2i}, x_{3i})$ — but the curse of dimensionality and difficulty of interpretation limit the utility of such models.

7. Generalized Nonparametric Regression

Generalized linear models encompass many of the statistical methods most commonly employed in data analysis, such as

- linear models with normally distributed errors
- logit and probit models for dichotomous response variables
- Poisson-regression (log-linear) models for counts.

A generalized linear models consists of three components:

1. A *random component*, in the form of a response variable y_i, which, conditional on the predictors, follows (in a traditional GLM) a distribution in an exponential family
   - normal
   - Poisson
   - binomial
   - gamma
   - inverse-normal
2. A *linear predictor*

$$\eta_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

   on which y_i depends.

3. A *link function* $g(\cdot)$ that transforms the mean of the dependent variable $\mu_i = E(y_i)$ to the linear predictor η_i. Standard link functions include:
   - The identity link: $g(\mu_i) = \mu_i$;
   - the log link: $g(\mu_i) = \log_e \mu_i$;
   - the inverse link: $g(\mu_i) = 1/\mu_i$;
   - the square-root link: $g(\mu_i) = \sqrt{\mu_i}$;
   - the logit link: $g(\mu_i) = \log_e \dfrac{\mu_i[}{1 - \mu_i}$;
   - the probit link: $g(\mu_i) = \Phi(\mu_i)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution; and
   - the complementary log-log link: $g(\mu_i) = \log_e [-\log_e(1 - \mu_i)]$.

- The logit, probit, and complementary log-log links are intended for binomial data, where y_i represents the observed proportion of 'successes' in n_i binomial trials, and μ_i is the probability of successes.
- In many applications, all of the n_i are 1, in which case y_i is either 0 or 1; this case is described as *binary* data.
- The logit and probit links are very similar; in particular, both approach $\mu = 0$ and $\mu = 1$ asymptotically and symmetrically.
- The complementary log-log link is asymmetric and may therefore be appropriate in a generalized linear model when the logit and probit links are not.
- In generalized nonparametric regression, the regression curve is flexible, and either the logit or probit link could be used to model an asymmetric approach to 0 and 1.
- As long as a generally reasonable link function is employed, the specific choice of link is not crucial in nonparametric regression.

Generalized nonparametric regression models retain the random component and link function of the generalized linear model, but substitute a smooth function of the x's for the linear predictor:

$$\eta_i = f(x_{1i}, x_{2i}, ..., x_{ki})$$

Likewise, *generalized additive models* express the transformed expectation of y as a sum of smooth functions of several predictors:

$$\eta_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki})$$

7.1 Local Likelihood Estimation

Figure 29 demonstrates why generalized regression models are needed (and why scatterplot smoothing is especially helpful for dichotomous responses).

The data on which this figure is based are from the 1976 U. S. Panel Study of Income Dynamics, and were originally analyzed by Mroz (1987); the data were employed by Berndt (1991) in an exercise on linear logistic regression, and by Long (1997) to illustrate this method.

- The response variable is married women's labour-force participation, with 'yes' coded as 1 and 'no' as 0.
- The predictor is the log of the woman's estimated wage rate.
  - The estimated wage is the actual wage rate for women who are in the labour force.
  - For women who are not in the labour force, the wage rate is estimated on the basis of a preliminary regression.

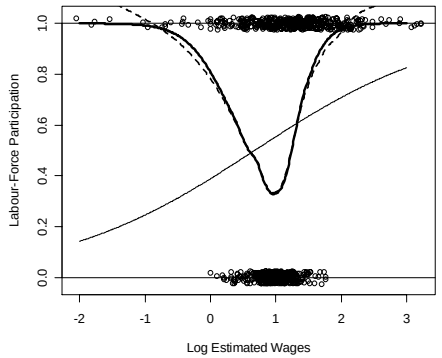

Figure 29. Scatterplot of labor-force participation (1 = Yes, 0 = No) by the log of estimated wages. The points are vertically jittered. Fits: linear logistic regression (light solid line); local linear-logistic regression (heavy solid line); local linear-least-squares regression (broken line).

- Log wages are much less variable for women who are not in the labour force: Predicted values are *expected* to be less variable than observed values.
- The points are "jittered" vertically, but the summary curves on the graph are fit to the unjittered data:
  (a) The linear logistic regression suggests a positive relationship between labour-force participation and estimated wages. Fitted values between 0 and 1 are interpretable as the estimated proportion of women in the labour force at various wage levels.
  (b) The local linear least-squares fit suggests a curvlinear relationship. The fit is defective in that it produces fitted values larger than 1 at the extremes of estimated wages. Moreover assumptions of constant error variance and normal errors are insupportable for binary data.
  (c) The local linear-logistic regression (to be described presently) is similar to the local least-squares regression except when the fitted proportion gets close to 1.

- The curvilinear pattern of the regression function is probably an artifact of the construction of estimated wages: Because estimated wages are less variable for women not in the labour force, more extreme values are observed for those in the labour force.

Generalized linear models are typically estimated by the method of maximum likelihood.

- The log-likelihood for these models takes the general form

$$\log_e l = \sum_{i=1}^{n} l(\mu_i; y_i)$$

where the y_i are the observed values of the response variable, and

$$\mu_i = E(y_i) = g^{-1}(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})$$

g^{-1} is the inverse of the link function (called the *mean function*).

- For example, for a binary logistic-regression model, the components of the log-likelihood are

$$l(\mu_i; y_i) = y_i \log_e \mu_i + (1 - y_i) \log_e (1 - \mu_i)$$

and the expected value of y is

$$\begin{aligned} \mu_i &= g^{-1}(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \\ &= \frac{1}{1 + \exp[-(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})]} \end{aligned}$$

- The maximum-likelihood estimates of the parameters are the values $\widehat{\alpha}, \widehat{\beta}_1, ..., \widehat{\beta}_k$ that maximize $\log_e l$.

In generalized nonparametric regression, we estimate the regression function at some set of focal values of the predictors.

- For simplicity, suppose that there is one predictor x, that the response variable is dichotomous, and that we want to estimate $\mu|x$ at the focal value x_0.
- We can perform a logistic polynomial regression of the form

$$\log_e \frac{\mu_i}{1-\mu_i} = \alpha + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + \cdots + \beta_p(x_i - x_0)^p$$

  maximizing the weighted log-likelihood

$$\log_e l_w = \sum_{i=1}^{n} w_i l(\mu_i; y_i)$$

  where $w_i = K[(x_i - x_0)/h]$ are kernel weights. Then $\widehat{\mu}|x_0 = g^{-1}(\widehat{\alpha})$.

- To trace the estimated regression curve, as in Figure 29, we repeat this procedure for representative values of x or at the observed x_i.
- As in local linear least-squares regression, the window half-width h can either be fixed, or can be adjusted to include a fixed number of nearest neighbors of the focal x.
- The extension of this approach to multiple regression is straightforward, but the curse of dimensionality and the difficulty of interpreting higher-dimensional fits are no less a problem than in local least-squares regression.

7.2 Generalized Additive Models

The generalized additive model replaces the parametric terms in the generalized linear model with smooth terms in the predictors:

$$\eta_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki})$$

- Local likelihood cannot be easily adapted to estimating the generalized additive model.
- An alternative is to adapt the method of *iteratively reweighted least squares (IRLS)*, which is typically used to obtain maximum-likelihood estimates for generalized linear models, combining it with back-fitting (or another suitable approach).

7.3 Statistical Inference

Once again, I will concentrate on binary logistic regression, with similar results applying to other generalized additive models. Similar results also apply to generalized nonparametric regression models estimated by local likelihood.

7.3.1 Confidence Envelopes*

- After the IRLS-backfitting procedure converges, the fitted values $\widehat{\eta}_i$ can be written as a linear transformation of quantities z_j called "pseudo-response" values (which are used in the IRLS fit),

$$\widehat{\eta}_i = r_{i1}z_1 + r_{i2}z_2 + \cdots + r_{in}z_n = \sum_{j=1}^{n} r_{ij}z_j$$

- The pseudo-response z_j has estimated asymptotic variance $1/[\widehat{\mu}_j(1-\widehat{\mu}_j)]$, and because the observations are asymptotically independent, the estimated asymptotic variance of $\widehat{\eta}_i$ is

$$\widehat{\mathcal{V}}(\widehat{\eta}_i) = \sum_{j=1}^{n} \frac{r_{ij}^2}{\widehat{\mu}_j(1-\widehat{\mu}_j)}$$

- An approximate pointwise 95-percent confidence band for the fitted regression surface follows as

$$\widehat{\eta}_i \pm 2\sqrt{\widehat{\mathcal{V}}(\widehat{\eta}_i)}$$

- The endpoints of the confidence band can be transformed to the probability scale by using $\mu = 1/[1+e^{-\eta})]$.
- Approximate confidence bands can also be constructed for the individual partial-regression functions, f_j.

7.4 Hypothesis Tests

Likelihood-ratio tests of hypotheses for generalized linear models are typically formulated in terms of the *deviance* for alternative, nested models.

- The deviance for a model is the log-likelihood-ratio statistic contrasting the model with a maximally specified or 'saturated' model, which dedicates a parameter to each observation.
  - Let $\ell(\boldsymbol{\mu};\mathbf{y})$ represent the log-likelihood for the model in question, and $\ell(\mathbf{y};\mathbf{y})$ the log-likelihood for the saturated model.
  - The deviance is then

$$D(\boldsymbol{\mu};\mathbf{y}) = -2[\ell(\boldsymbol{\mu};\mathbf{y}) - \ell(\mathbf{y};\mathbf{y})]$$

  - The deviance is a generalization of the residual sum of squares for a general linear model with normal errors.

- For the binary linear-logit model, the log-likelihood for the saturated model is 0, leading to a particularly simple expression for the deviance:

$$\begin{aligned} D(\boldsymbol{\mu};\mathbf{y}) &= -2\ell(\boldsymbol{\mu};\mathbf{y}) \\ &= -2\sum_{i=1}^{n}\ell(\mu_i;y_i) \\ &= -2\sum_{i=1}^{n}[y_i \log_e \widehat{\mu}_i + (1-y_i)\log_e(1-\widehat{\mu}_i)] \end{aligned}$$

- The degrees of freedom associated with the deviance are n minus the number of parameters in the model.

- The log likelihood-ratio statistic for contrasting nested models is the difference in the deviances for the two models. This statistic is asymptotically distributed as chi-squared, with degrees of freedom given by the difference in the number of parameters for the models.
- Tests based on the deviance can be extended by analogy to generalized nonparametric-regression models, including generalized additive models. Degrees of freedom can be defined analogously to nonparametric regression models, as described previously.

7.5 An Illustration: Mroz's Labour-Force Data

I will adapt an example that appears in Long (1997), based on Mroz's married women's labour-force participation data, substituting a semiparametric logistic regression for the linear logistic regression in Long.

- The response variable is binary: labour-force participation (`lfp`).
- The predictor variables are as follows:

Predictor	*Description*
`k5`	number of children ages 5 and younger
`k618`	number of children ages 6 to 18
`age`	women's age in years
`wc`	wife's college attendance
`hc`	husband's college attendance
`inc`	family income excluding wife's income

Predictor	*Remarks*
`k5`	0–3, few 3's
`k618`	0–8, few >5
`age`	30–60, single years
`wc`	0/1
`hc`	0/1
`inc`	\$1000s

- Because `k5` and `k618` are discrete predictors with small numbers of distinct values, I modeled these terms as sets of dummy regressors, capable of capturing any form of partial relationship to labour-force participation.
- `wc` and `hc` are also dummy regressors, representing dichotomous predictors.
- `age` and `inc` are modeled using nearest-neighbour locally linear logistic smoothers; in each case, visual trial and error suggested a span of .5.

Figure 30 graphs the estimated partial-regression functions for the semiparametric logistic regression fit to Mroz's data (model 1 in the table below).

- Each panel of the figure shows the partial-regression function for one of the predictors, along with a pointwise 95-percent confidence envelope and partial residuals.
- The vertical (`lfp`) axis of each plot is on the logit scale.
- Panels (a) and (b), for age and family income, show the fit of local partial logistic regressions. The nonlinearity in each of these partial regressions appears slight; we will determine presently whether the departure from linearity is statistically significant.
- The partial regression for children five and under, in panel (c), also appears nearly linear.

- labour-force participation seems almost unrelated to number of children six to 18, in panel (d)
- labour-force participation appears to rise with wife's college attendance (panel e), and is apparently unrelated to husband's college attendance (panel f).
- Note: For the plots, I used versions of `k5` with the last category as "2 or more," and of `k618` with the last category as "5 or more."

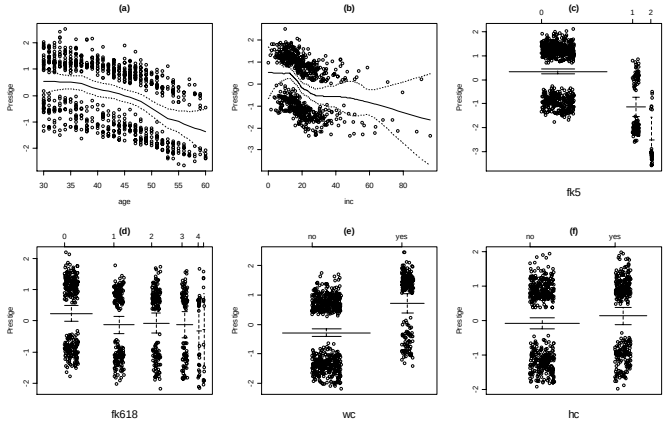

Figure 30. Estimated partial-regression functions for the semi-parametric logistic regression of married women's labor-force participation on several predictors.

- Here are the deviance and residual degrees of freedom for several models fit to Mroz's data (using the `gam` function in the R `gam` package, correcting the degrees of freedom):

	Predictors							
Model	`age`	`inc`	`k5`	`k618`	`wc`	`hc`	*Deviance*	df_{res}
0	L	L	L	L	D	D	922.27	746.0
1	S	S	D	D	D	D	906.40	730.6
2	L	S	D	D	D	D	910.04	733.2
3	S	L	D	D	D	D	913.10	734.4
4	S	S	L	D	D	D	907.12	732.6
5	S	S	D	L	D	D	911.78	737.6

The following code is used for terms in the models:

– L, a linear term;

– D, a dummy regressor or set of dummy regressors;

– S, a local-linear logit smooth.

- From these models, I calculated the following tests for nonlinearity, all of which are nonsignificant:

Predictor	*Models Contrasted*	*Difference in Deviance*	*df*	p
`age`	$2-1$	3.64	2.6	.24
`inc`	$3-1$	6.70	3.8	.14
`k5`	$4-1$	0.72	2.0	.70
`k618`	$5-1$	5.38	7.0	.61

- An overall test, contrasting the linear-logit model (0) with the semiparametric logit model (1), produces a difference in deviance of 15.62 on 15.5 degrees of freedom, which is also nonsignificant ($p = .44$).

- To test for each of the terms in the semiparametric model, I fit the additional models:

	Predictors							
Model	`age`	`inc`	`k5`	`k618`	`wc`	`hc`	*Deviance*	df_{res}
6	–	S	D	D	D	D	935.13	734.2
7	S	–	D	D	D	D	930.16	735.4
8	S	S	–	D	D	D	968.30	733.5
9	S	S	D	–	D	D	913.98	738.6
10	S	S	D	D	–	D	927.03	731.6
11	S	S	D	D	D	–	907.50	731.6

- The analysis of deviance table is as follows:

Predictor	*Models Contrasted*	*Difference in Deviance*	*df*	p
`age`	$6-1$	28.73	3.6	$< .00001$
`inc`	$7-1$	23.76	4.8	$.0002$
`k5`	$8-1$	61.90	2.9	$\ll .00001$
`k618`	$9-1$	7.58	8.0	$.48$
`wc`	$10-1$	20.63	1.0	$< .00001$
`hc`	$11-1$	1.10	1.0	$.29$

- There is strong evidence of partial relationships of women's labour-force participation to age, family income, children five and under, and wife's college attendance, but not to children six to 18 or to husband's college attendance.

